Fostering Serendipitous Knowledge Discovery using an Adaptive Multigraph-based Faceted Browser

Ali Khalili
Department of Computer Science
Vrije Universiteit Amsterdam
the Netherlands
a.khalili@vu.nl

Pek van Andel
The University Medical Center Groningen
the Netherlands
m.v.van.andel@umcg.nl

Peter van den Besselaar
Department of Organization Sciences
Vrije Universiteit Amsterdm
the Netherlands
p.a.a.vanden.besselaar@vu.nl

Klaas Andries de Graaf
Department of Computer Science
Vrije Universiteit Amsterdam
the Netherlands
ka.de.graaf@vu.nl

ABSTRACT

Serendipity, the art of making an unsought finding plays also an important role in the emerging field of data science, allowing the discovery of interesting and valuable facts not initially sought for. Previous research has extracted many serendipity-fostering patterns applicable to digital data-driven systems. Linked Open Data (LOD) on the Web which is powered by the Follow-Your-Nose effect, provides already a rich source for serendipity. The serendipity most often takes place when browsing data. Therefore, flexible and intuitive browsing user interfaces which support serendipity triggers such as enigmas, anomalies and novelties, can increase the likelihood of serendipity on LOD. In this work, we propose a set of serendipity-fostering design features supported by an adaptive multigraph-based faceted browsing interface to catalyze serendipity on Semantic Web and LOD environments.

1 INTRODUCTION

“Unless you expect the unexpected you will never find [truth], for it is hard to discover and hard to attain.” -*Heraclitus*[1]

The experience of ‘accidental’ discovery and acquisition of information generally known as *Serendipity* refers to ‘accidentally’ bumping into (new, true, useful, or personal interest-related) information, initially not looked for. Serendipity, defined as the art of making an unsought finding[33], has played a pivotal role in the discovery of many drugs. Major types of psychotropic drugs (effecting mental activity and behavior) such as *Lithium*, *Chlorpromazine* and *Imipramine* were serendipitously discovered in the 1950s and 1960s. In 2012, [11] reported that 24% of all pharmaceuticals on the market and in particular 35.2% of all the anticancer drugs in clinical use were discovered by serendipity.

Serendipity also plays an important role in the emerging field of data science by enhancing information retrieval[14] and by promoting unexpected knowledge discovery. The World Wide Web has provided a global information space comprising billions of connected documents. “The unexpected connection is more powerful than one that is obvious”, as aptly asserted by Haraclitus in 500 BC. However, most of the existing centralized “nearest neighbor” search approaches on the Web, such as Google, although very useful in finding explicitly relevant results, are killing serendipity by excessively limiting the encountering of unexpected information[1]. On the other hand, the ever-growing amount of Linked Data publicly accessible and distributed on the Web increases the likelihood that some of the data, which will make an impact in our professional or private lives will come to us by chance—without searching it initially. The adoption of Semantic Web as a linked information space in which data are dynamically enriched and added, provides an open interactive system, with external links and the ability to make information easily accessible, re-usable including the possibility of the discovery and serendipitous reuse of other related information[2, 29].

‘Unsought discoveries’ most often take place in the context of browsing unbounded data spaces; people immerse themselves in the items that interest them, meandering from topic to topic, and so on and so forth (i.e., the *Follow-Your-Nose* method[35] to traverse the given semantic links from a resource) while concurrently remarking interesting and informative information en route[32]. Therefore, flexible and intuitive browsing user interfaces (UIs) which support serendipity triggers, can increase the likelihood of accidental knowledge discovery on Linked Open Data (LOD). Although there has been some research on supporting serendipity through query modifications and semantic path-finding on knowledge graphs, we still lack UIs that increase the emergence of serendipities on LOD. In this paper, we aim to provide an adaptive multigraph-based faceted browsing interface to foster serendipity on Semantic Web and LOD environments. The contributions of this work are in particular:

- Proposing a set of serendipity-fostering design features which are applicable to data-driven environments, by conducting an extensive literature review.
- Presenting a set of UI and Semantic Web-based techniques to support the proposed serendipity design features in the context of Linked Open Data.
- Implementing an open-source adaptive multigraph-based faceted browser to facilitate serendipity while browsing linked data.
- Discussing a set of in-use cases to demonstrate the capabilities of our implemented solution.

[1] according to secondary sources

2 SERENDIPITY: THE ART OF UNSOUGHT FINDING

The word "serendipity" was coined in 1754 by Horace Walpole, a letter writer and politician[23]. Walpole was inspired by an old Persian fairy tale known as "The Three Princes of Serendip" published in 1302 AD by Amir Khusrow Dehlavi[2]. The original story is about three princes from Serendip (a medieval Persian name for Sri Lanka), well trained in the art of tracking, who make ten 'accidental' discoveries via ten surprising observations, and by interpreting all ten correctly, on their grand tour to see the different countries and miracles of the world. Walpole created the word serendipity to refer to "always making discoveries, by accidents and sagacity, of things they (the three princes of Serendip) were not in quest of" or "a surprising observation followed by a correct hypothesis".

In our view, serendipity consists of two main steps: a surprising observation (*trigger*) and then a correct interpretation (*abduction*). The trigger is a riddle, an anomaly, or a novelty. Abduction[34] refers to the process of guessing, interpreting, creating and testing hypotheses in order to find a correct explanation, one that is evidence-based. As stated in [27], you do not reach Serendip by plotting a course for it. You have to set out in good faith for elsewhere and lose your bearings...serendipitously!

Serendipitous discovery may be facilitated but it is by definition an emergent process[7]. Because it is an emergent process, transitioning serendipity to a science where certain patterns are defined is an inherently difficult task to be managed. What we can offer is to foster the process of serendipity by providing an incubator-like environment for serendipity. In other words, the environment will increase the likelihood of serendipity, without guaranteeing it. In the context of a knowledge building environment, [1] calls such a system that supports both *trigger* and *abduction* an "inspiration engine" or in [11], the term "pseudo-serendipity" is used to describe the approach of such systems i.e. *a sought finding, found on an unsought road.* As result of our extensive literature review and consultation with a serendipitiologist, we extracted, blended and adapted a set of serendipity-fostering design features that are applicable to data-driven systems. The main sources of inspiration for these features come from [4, 6, 32, 33]. These features can coexist, overlap, cooperate, complement and reinforce each other. In the following sections, we describe these 12 serendipity-fostering design features together with some ideas how to realize them:

2.1 Design Features Related to Observations

F_1: Make surprising observations more noticeable.
Surprising observations are the main initiators of accidental knowledge discovery. A single surprising observation, especially if it is repeatedly done, or multiple different surprising observations, when they refer to the same phenomenon, can trigger serendipity. Creative data visualization is an activity that enables users to make hypotheses, look for patterns and exceptions, and then refine their hypothesis. Users might find surprising results that shake their established beliefs, provoke new insights, and possibly lead to important discoveries[30]. Users often need to look at the same data from different perspectives. Therefore, tools that provide different views on data can foster serendipity.

F_2: Make errors in data more visible in order to detect successful errors easier.
Errors and exceptions are not always accidental and can sometimes indicate the real and natural behavior of a system known as "desire lines"[24]. Following the trails left behind quantitative and qualitative anomalies in data can result in new insights. Semantic Web tools, such as Shapes Constraint Language (SHACL)[3], or restrictions supported by RDF-S & Web Ontology Language (OWL), which allow validating RDF graphs against a set of rules and conditions, help to automate the discovery of successful errors and thereby facilitate serendipity.

F_3: Allow inversion and contrast.
The inversion and contrast features depict the unexpected aspect of serendipity. Sometimes turning things upside down or inside out allows us to watch those things from another perspective and to discover gaps in knowledge. Looking at the insights in the opposite direction than intended by users will and can cause to a breakthrough discovery. This feature can be supported by SPARQL query inversion where a query is adapted to include results which were not returned by the initial query; or the query employs RDF properties which contrast with the initial properties used.

F_4: Support randomization and disturbance.
Chance can be used intentionally in serendipitous knowledge discovery. Randomization and disturbance are two methods to increase the chance encounter. For example, the Randomised Coffee Trial (RCT), is a technique used by some firms to create an institutionalized space for serendipity through connecting people in the firm at random and give them time to meet to have a coffee and talk about whatever they wish. In a linked data browsing system, randomizing the items (or modifying the order of the sets of triples) presented on top of the result lists can increase the probability of the chance encounter. It also serves as an efficient solution to the problem of 'blind spots' and to decrease the possibility of bias in interpreting results.

F_5: Allow monitoring of side-effects when interacting with data.
Accidental discoveries through observation of side-effects has played a crucial role in drug/treatment discovery. For example, *Dimenhydrinate* was first developed as an antihistamine, but is now sold as a travel sickness medication owing to a surprising observation/realization by one of the participants in the clinical trials[11]. A system that consistently monitors the side-effects of user interactions with data and provides appropriate feedback on surprising observations implied, can facilitate serendipity.

F_6: Support detection and investigation of by-products.
Some serendipitous discoveries have occurred as by-products or spin-offs of the main product which was intended to come out. A user searches for A and, as a by-product, finds B as a surprising unsought result. A system that supports detection and investigation of by-products resulted from user interactions can foster serendipity. Error, enigma, anomaly and novelty detection mechanisms suggested by F_4 can support this feature as well.

F_7: Support background knowledge and user contextualization.
"In the sciences of observation, chance favors only prepared minds",

[2] https://en.wikipedia.org/wiki/Amir_Khusrow

[3] https://www.w3.org/TR/shacl

said Louis Pasteur. Most of the serendipitous discoveries are triggered by chance or a chance encounter. A chance encounter occurs at the point in human interaction with an information system when a human makes an accidental discovery. The encounter is generally influenced by the person's prior knowledge, although not necessarily, and by the person's recognition of the affordances. A serendipity-fostering environment depends both on the information seeker and the medium. Without basic topical knowledge, there is no capacity to observe and interpret the surprising facts correctly[14]. Techniques for integrating user profiles and domain knowledge into query processing[31] can improve the relevancy of the query results obtained by users and thereby promote the serendipity.

F_8: Support both convergent and divergent information behavior. When users move through an information space they may change directions and behavior several times as their information needs and interests develop or get triggered depending on affordances encountered on their way through the information space. Supporting both convergent and divergent information behavior[4] in a data-driven system facilitates serendipity. Convergent (depth first, focused, not easily distracted) behavior is supported by features that allow zooming in and narrowing the vision of users while divergent (breadth first, creative, but easily distractible) behavior is supported by features that allow zooming out and broadening the vision of users.

2.2 Design Features Related to Explanation of the Observations

F_9: Facilitate the explanation of surprising observations. After the occurrence of a surprising observation as trigger, abduction is needed to understand why and how this accidental event is entailed. Abduction gives some clues to interpret the surprising result and to find the correct explanation for it. Metadata and provenance as means to support causal reasoning aid to provide reasons and explanation for surprising observations. Provenance also helps to assess the quality, reliability, or trustworthiness of surprising data which is discovered. Exploiting existing provenance data models and ontologies[4] on the Semantic Web can foster serendipity.

F_{10}: Allow sharing of surprising observations among multiple users. A surprising observation done by user A, when correctly explained by user B, can result in positive serendipity. Tools such as *YASGUI*[5], *grlc*[6] and *BASIL*[7] help to support this feature via sharing and modification of SPARQL queries among multiple users through a standardized Web API.

F_{11}: Enable reasoning by analogy. Analogical learning as the act of finding similar entities or phenomena when studying an entity or phenomenon has been long known as an approach for knowledge transfer. Analogical reasoning can happen either in the same or a completely different context than the original context of data. Semantic Web-based knowledge abstraction techniques on LOD help to foster serendipity by enabling the abstraction of the knowledge representation structure related to a particular knowledge artifact, by analyzing its constituent elements and their relationships. For instance, by employing SPARQL query patterns one can identify similar resources to a resource of interest by considering resources with similar RDF properties and values or with more generalized RDF classes than the resource of interest. With regards to analogical reasoning on different context (e.g. concepts from business domain which are similar to concepts in medical domain), there are less strategies discussed in the literature. In [20], a framework is proposed for explicitly modeling analogical structures in multi-relational or knowledge graph embedding. Another possible strategy is to analyze ontology design patterns instead of concrete entity-similarity metrics to represent relevance between entities in one context to entities in another context[5].

F_{12}: Support extending the memory of user by invoking provocative reminders and relevance feedback. Keeping track of previous user interactions, queries, and resulting data while browsing complex data enables a data-driven system to invoke provocative messages as reminders to help extend the memory of users when interacting with other related datasets. When potentially valuable information is encountered an important ability would be the capacity to recognize it and its "affordances"[25]–clues about how it can be used. *Relevance feedback*[22]–asking information seekers to make relevance judgments about returned objects and then executing a revised query based on those judgmentsâĂŤ-is already known as a powerful way to cultivate knowledge discovery. If a person is not alert enough, the message remains unnoticed regardless of its potential value. A system that provides users with meaningful reminders connected to their past browsing experience can increase the likelihood of serendipity.

3 AN ADAPTIVE MULTIGRAPH-BASED FACETED BROWSER: A TRIGGER AND FACILITATOR FOR SERENDIPITY

There are generally three ways in which people discover and acquire information: 1) The *Purposive search*: A directed search looking for a definite piece of information. 2) *Exploratory search and browsing*: A general purpose semi-directed search and browsing of data deliberately looking for an object that cannot be fully described or to get inspiration by looking at some items of interest. 3) *Capricious search and browsing*: An undirected random search and browsing of information without a defined goal. Accidental knowledge discoveries occur most frequently during this type of unplanned investigative search and browsing.

Systems that support the first, prompt users for search terms and keywords, and provide options for *parametric search* allowing users to manipulate queries and results by visually specifying a set of constraints. The focus of such systems which are well supported by current Web search engines is on *precision* i.e. minimizing the number of possibly irrelevant objects that are retrieved.

Hypermedia, menu-driven and faceted navigation systems that provide views and overviews of the data facilitate the second. Faceted navigation fills in the piece that is missing in parametric search: *guidance*. Parametric search requires that the user express an information need as a query in one shot, making selections

[4] https://www.w3.org/TR/prov-overview
[5] http://yasgui.org
[6] http://grlc.io
[7] http://basil.kmi.open.ac.uk

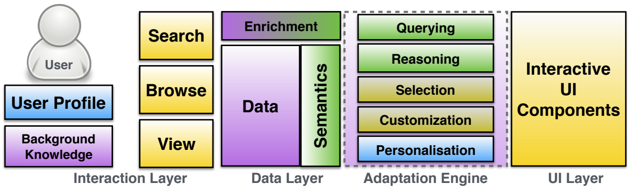


Figure 1: Architecture of the proposed adaptive faceted browsing environment.

across all facets of interest. In contrast, *faceted navigation* allows the user to elaborate a query progressively, seeing the effect of each choice in one facet on the available choices in other facets. Systems that support this type of browsing are more concerned with recall i.e. maximizing the number of possibly relevant objects that are retrieved.

The third type, the serendipitous approach, is a type of information seeking that is not traditionally examined in information retrieval research and has received little attention by both developers and researchers. In this paper we focus on this latter type by augmenting the existing faceted browsing techniques with serendipity-fostering features discussed in Section 2. We call our proposed adaptive faceted browsing environment “FERASAT”[8] (FacEted bRowser And Serendipity cATalyzer). FERASAT is built on top of the LD-R framework[19] to enable skeuomorphic, adaptive and component-based design of the system. Skeuomorphism[25] in UI design is employed to incorporate recognizable UI elements which are familiar to users and thereby decrease the cognitive load of users when interacting with the system. Skeuomorphic design in FERASAT is a way to bypass the *Pathetic Fallacy of RDF*[9] [17].

Figure 1 depicts the architecture of the FERASAT where related elements are color coded. The system provides three main modes of interaction with data namely search, browse and view. During the user interactions, based on the semantics of data and the given user context, the system adapts its behavior by rendering appropriate interactive UI components. In the following sections we describe the main building blocks of the FERASAT environment together with how they support serendipity-fostering features discussed in Section 2:

3.1 Interaction Layer

According to the theory of “Seven Stages of Action”[25] which explains the psychology of a person behind a task, user interactions with a system occur in two gulfs namely a *gulf of execution* and a *gulf of evaluation*. The gulf of execution focuses on allowable interactions (i.e. affordances) in the system, whereas the gulf of evaluation reflects the amount of effort that the person must exert to interpret the state of the system after an interaction. Within the FERASAT environment, interactions in the gulf of execution (e.g. inverting a selected facet) are used as triggers for serendipity and interactions in the gulf of evaluation (e.g. visualization & in-detail browsing of the properties of a resource) are used to support the process of abduction.

Figure 2 shows a mock-up of the design we devised for the FERASAT faceted browser. When browsing a set of linked data

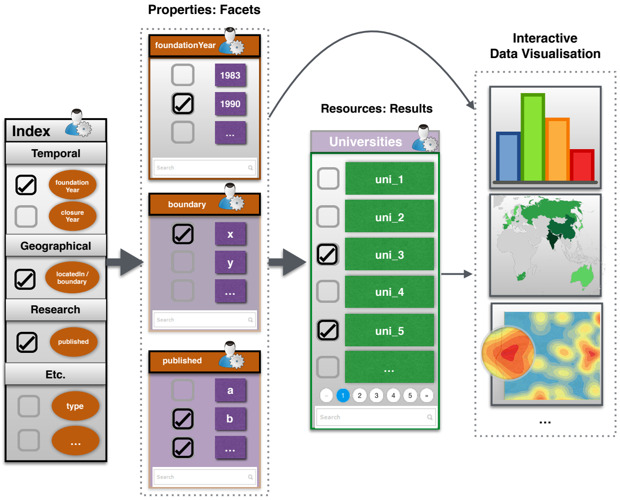


Figure 2: A mock-up of the adaptive faceted browser.

which is scattered over multiple knowledge graphs (e.g. Figure 3), the first step is to identify properties of interest as semantic links to move forward and backward in the data space. The index facet lists these designated RDF properties grouped by the aspect they are addressing. According to [26], the representative properties should provide best descriptors and navigators for the underlying knowledge graphs to be browsed. There can be also derivative and new properties dynamically added to the environment as the user proceeds with browsing data.

In the initial state, all the RDF resources are displayed without any constraints. When a user selects a property, a new facet is generated to display the object values of the selected property together with the number of resources containing those values. The facet can be configured to employ different interactive UI components (e.g. charts, maps, etc.) to render the values of a selected property. Flexible UI components support the features F_1, F_2 and F_9 by allowing users to exploit multiple interactive visualizations to do surprising observations and also discover successful errors in data together with the possible explanations for their occurrence. The list of values in a facet can be shuffled to change the ordering based on a random factor or some criteria other than the default sorting criterion which is the frequency of the corresponding resources (supporting F_4). When a user selects one or more values of an active facet, a SPARQL query with the corresponding constraints is generated and executed to update the results list. Users can invert the selected values in a facet to see the results which exclude those selected values (supporting F_3). If multiple facets are active, any change to a facet will affect the remaining active facets to take into account the constraint imposed by that change (supporting F_5). Users can focus on each facet, search within its values and view in-detail characteristics of each object value (supporting both convergent and divergent information behavior presented by F_8). When a user is browsing a facet which was browsed before, the UI provides some reminders as pop-ups about the previous usage of that facet (supporting F_{12}).

[8] in Persian, the term 'Ferasat' refers to the ability of intuitive knowledge acquisition.

[9] i.e. display RDF data to the users as a graph because the underlying data model is a graph.

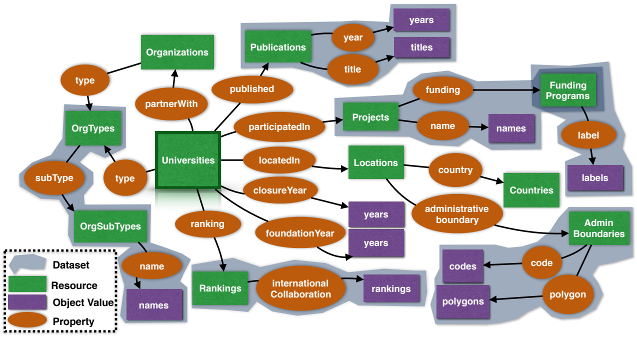


Figure 3: An example of the Linked Data scattered over multiple graphs (datasets).

The result list is the terminal facet in the system which shows the final result of the generated SPARQL query as a set of RDF resources constrained by the selected RDF properties and values. Clicking on a resource reveals the detailed characteristics of that resource (supporting F_9 to study a particular surprising observation). To further investigate the results, a user can select multiple resources and ask for the potential correlations between them (supporting F_6 to investigate by-products). For example, given the linked data in Figure 3, a user might want to browse and find the relation between entities of type universities which are founded in certain years AND are located in certain administrative boundaries AND have published on certain research topics. Using the faceted browser, such a query is generated in a progressive way where users can investigate the effect of each selected facet on other facets and on the results list, while traversing multiple distributed knowledge graphs (in this case, graphs that provide data about universities connected to graphs that provide values related to publications and administrative boundaries), until the full query is answered.

3.2 Data Layer

There are five different sorts of data taken into account within the FERASAT environment: 1) *user's profile data* to understand the user preferences, 2) *user's background knowledge* to consider a user's domain of interest while browsing data, 3) *original data* to be browsed, 4) *configuration data* as output of adaptation process to customize and personalize both data and UIs, 5) *complementary data* added as enrichment to original data for richer contextualization. All the above datasets are represented as single or multiple RDF graphs (e.g., Figure 3) to be ready for integration (using federated SPARQL queries) and analysis.

FERASAT supports resource annotation to interlink the original data with the user's background knowledge and to generate complementary data connected to the original data to be browsed (supporting F_7 for user contextualization by giving users additional contextual facets to complement their browsing experience). At the moment, two types of annotation are supported within the system: *Named Entity Recognition* (NER) using DBpedia Spotlight[10] and *Geo-boundary-tagging* supported by open geo boundaries from OpenStreetMap and GADM[11]. There are interactive UIs embedded in the FERASAT system to interactively annotate a dataset before the browsing activity starts.

3.3 UI Layer

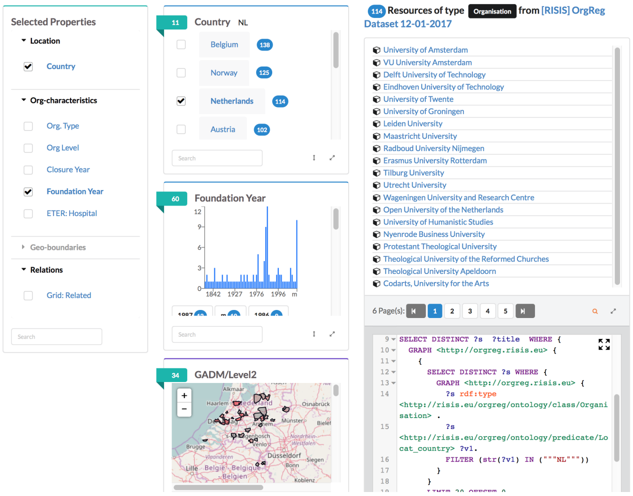
FERASAT exploits flexible and interactive UI components brought by the LD-R UI framework[19]. There are four core component levels in an LD-R Web application. Each core component abstracts the actions required for retrieving and updating the graph-based data and provides a basis for user-defined components to interact with Linked Data in three modes: search, browse and view. The data-flow in the system starts from the Dataset component which handles all the events related to a set of resources under a named graph identified by a URI. The next level is the Resource component which is identified by a URI and indicates the RDF resource to be described in the application. A resource includes a set of properties which are handled by the Property component. Properties can be either individual or aggregate when combining multiple features of a resource (e.g. a component that combines longitude and latitude properties; start date and end date properties for a date range, etc.). Each property is instantiated by an individual value or multiple values in case of an aggregate object. The value(s) of properties are controlled by the Value component. In turn, Value components invoke different components to search, browse and view the property values. Value components are terminals in the LD-R single directional data flow where customized user-generated components (e.g. charts, maps, diagrams, etc.) can be plugged into the system.

3.4 Adaptation Engine

An adaptive UI[12] is a UI which adapts, that is, changes its layout and elements to the needs of the user or context and is similarly alterable by each user. In the context of FERASAT, we devise a particular type of adaptive UI called a *data-aware UI* [18] that a) can understand users' data and b) can interact with users accordingly. As depicted in Figure 1, FERASAT incorporates an adaptation engine to realize data-aware UIs when users interact with data. The task of adaptation engine is to make a bridge between data (enriched by semantics) and existing UI components suitable to render data. The adaptation engine includes the following core components:

- *Querying*. This part is responsible for composing, sharing and running of SPARQL queries within the FERASAT environment. FERASAT exploits *YASGUI* and *grlc* to allow identifying, sharing and repurposing of SPARQL queries among multiple users (supporting F_{10}). It also provides a set of SPARQL query templates similar to the one discussed in [9] to find analogous resources within the same domain (partially supporting F_{11}).
- *Reasoning*. This is the core part of the engine where different datasets mentioned in subsection 3.2 are analyzed in an integrative way to find the best strategy for data rendering and UI augmentation.
- *Selection*. This part allows to manually or automatically, as result of reasoning select or replace and existing UI component.
- *Customization*. This part allows to manually or automatically customize an existing UI component.

[10] http://www.dbpedia-spotlight.org
[11] http://gadm.org
[12] http://en.wikipedia.org/wiki/Adaptive_user_interface



Selected Properties

- Location
  - ☑ Country
- Org-characteristics
  - ☐ Org. Type
  - ☐ Org Level
  - ☐ Closure Year
  - ☑ Foundation Year
  - ☐ ETER: Hospital
- Geo-boundaries
- Relations
  - ☐ Grid: Related

11 Country NL

☐	Belgium	138
☐	Norway	125
☑	Netherlands	114
☐	Austria	102

60 Foundation Year

114 Resources of type Organisation from [RISIS] OrgReg Dataset 12-01-2017

- University of Amsterdam
- VU University Amsterdam
- Delft University of Technology
- Eindhoven University of Technology
- University of Twente
- University of Groningen
- Leiden University
- Maastricht University
- Radboud University Nijmegen
- Erasmus University Rotterdam
- Tilburg University
- Utrecht University
- Wageningen University and Research Centre
- Open University of the Netherlands
- University of Humanistic Studies
- Nyenrode Business University
- Protestant Theological University
- Theological University of the Reformed Churches
- Theological University Apeldoorn
- Codarts, University for the Arts

6 Page(s): 1 2 3 4 5

```
SELECT DISTINCT ?s  ?title  WHERE {
  GRAPH <http://orgreg.risis.eu> {
    {
      SELECT DISTINCT ?s WHERE {
        GRAPH <http://orgreg.risis.eu> {
          ?s rdf:type
<http://risis.eu/orgreg/ontology/class/Organi
sation> .
          ?s
<http://risis.eu/orgreg/ontology/predicate/Lo
cat_country> ?v1.
          FILTER (str(?v1) IN ("""NL"""))
        }
      }
      LIMIT 20 OFFSET 0
```

Figure 4: An screenshot of the implemented adaptive faceted browser.

- *Personalization.* This part allows to manually or automatically personalize an existing UI component. Personalization will overwrite the configurations used for customization to consider the user's context.

The configuration process is done by traversing the hypergraph generated either manually by a user or automatically as result of reasoning. FERASAT exploits a hierarchical permutation of the Dataset, Resource, Property, and Value components as *scopes* to select specific parts of the UI to be customized or personalized, as described in [18, 19].

4 IMPLEMENTATION

FERASAT is implemented as a ReactJS component (backed by NodeJS) within the open-source Linked Data Reactor[13] (LD-R) framework and is available to download at http://ferasat.ld-r.org together with its documentation and demos (see Figure 4 for an screenshot of the FERASAT environment).

5 USE CASES

FERASAT is integrated into the SMS[14] (Semantically Mapping Science) platform as the technical core within the RISIS.eu project. It is actively used to browse data related to Science, Technology & Innovation (STI) studies. A complete list of use cases is available at http://sms.risis.eu/usecases. In this section, we provide a brief summary of two use cases related to serendipitous knowledge discovery in the STI domain written by two social scientists who experienced browsing data on FERASAT environment while conducting research:

1. Analyzing change in the research/Higher Education (HE) systems.

The RISIS datastore contains many datasets with information about organizations. I was mainly interested in structural change in HE systems by navigating through those datasets. The faceted browser was of great help, as it enabled me to explore the available information in a graphical form. While browsing the datasets, I found a property "foundation year". Selecting that property for a country, I got the frequency of new foundations of HE institutions per year (see Figure 4), and I saw immediately (F_1) a high concentration in two consecutive years: in 1986 and 1987 some 21 new HE institutions were founded in the Netherlands, on a total of 114: So some substantial changes in the HE system seem to have taken place! By selecting these two years, the list of organizations shows the names of the institutions that were founded in these two years. I could inspect the list, but also select a single institution and inspect the available information in the datastore, but also more broadly on the web, as all the organizations are also linked to their website and their Wikipedia page (F_7, F_8, F_9). So, I did not only have much numerical data in the data network, such as numbers of students and staff, but also qualitative (textual) data for further inspection. Looking at the various newly founded schools showed that these are all Universities of Applied Sciences, so the "second layer" Dutch HE institutions. By reading the historical information on their Websites, one would find out that the new founded institutions in fact are conglomerations of smaller schools into very large new institutions (F_7, F_9). This indeed can be considered as a major reform of the Dutch HE system.

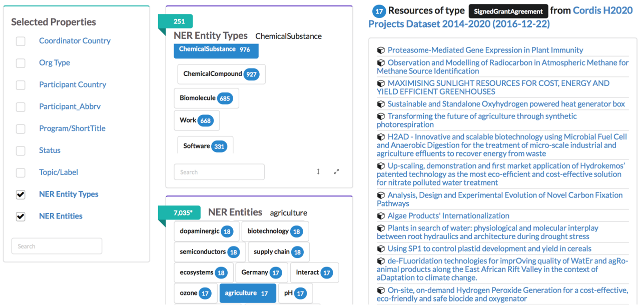
A follow-up question would be whether this is a typical Dutch phenomenon, or whether similar changes have taken place in other countries (F_{11}). Belgium could be a second case to inspect, and I followed the same steps. Indeed, as the browser shows, also here we find concentrations of foundations of new HE institutions, but now in the year 1995 when 32 new HE institutions were founded in Belgium. If I select the year 1995, I get a list with the names of the newly founded institutions and can further inspect the available information on those institutions. I did not have any prior knowledge on the Belgian system (F_7), but inspecting the list of names in the results, one immediately sees that the changes probably took place in the French speaking part of Belgium, as all institution names are French, and not in the Flemish speaking part (F_6). Indeed, the two language regions have their own HE system, so this could clearly be the case.

The third example I tried was Austria, and indeed also there I detected a concentration of new institutions in 2007 - a decade after the changes in Belgium and two decades after the changes in the Netherlands. Of the total of 102 HE institutions in Austria, 15 were created in 2007 - again a percentage suggesting some form of structural change. Even if one is completely unknowledgeable about the Austrian HE system, selecting the entity type in the browser tells us (F_5) that the changes have taken place in the sector of teacher education: the newly founded HE institutions are all of type "University of Education", 'University College of Teacher Education", and "Pedagogical University". Without further investigation, one already can conclude that the changes in the Austrian system are less broad than in the Netherlands or in Belgium, where the changes seem to cover a much larger part of the HE system.

2. Evaluating research portfolios with regards to current societal challenges.

I used the faceted browser to browse CORDIS open dataset on H2020 EU projects to evaluate research portfolios. The browser

[13] https://github.com/ali1k/ld-r
[14] http://sms.risis.eu



Selected Properties

□ Coordinator Country
□ Org Type
□ Participant Country
□ Participant_Abbrv
□ Program/ShortTitle
□ Status
□ Topic/Label
☑ NER Entity Types
☑ NER Entities

Search

251 NER Entity Types ChemicalSubstance

ChemicalSubstance	976
ChemicalCompound	927
Biomolecule	685
Work	668
Software	331

Search

7,035* NER Entities agriculture

dopaminergic	18
biotechnology	18
semiconductors	18
supply chain	18
ecosystems	18
Germany	17
interact	17
ozone	17
agriculture	17
pH	17

17 Resources of type SignedGrantAgreement from Cordis H2020 Projects Dataset 2014-2020 (2016-12-22)

- Proteasome-Mediated Gene Expression in Plant Immunity
- Observation and Modelling of Radiocarbon in Atmospheric Methane for Methane Source Identification
- MAXIMISING SUNLIGHT RESOURCES FOR COST, ENERGY AND YIELD EFFICIENT GREENHOUSES
- Sustainable and Standalone Oxyhydrogen powered heat generator box
- Transforming the future of agriculture through synthetic photorespiration
- H2AD - Innovative and scalable biotechnology using Microbial Fuel Cell and Anaerobic Digestion for the treatment of micro-scale industrial and agriculture effluents to recover energy from waste
- Up-scaling, demonstration and first market application of Hydrokemos' patented technology as the most eco-efficient and cost-effective solution for nitrate polluted water treatment
- Analysis, Design and Experimental Evolution of Novel Carbon Fixation Pathways
- Algae Products' Internationalization
- Plants in search of water: physiological and molecular interplay between root hydraulics and architecture during drought stress
- Using SP1 to control plastid development and yield in cereals
- de-FLuoridation technologies for imprOving quality of WatEr and agRo-animal products along the East African Rift Valley in the context of aDaptation to climate change.
- On-site, on-demand Hydrogen Peroxide Generation for a cost-effective, eco-friendly and safe biocide and oxygenator

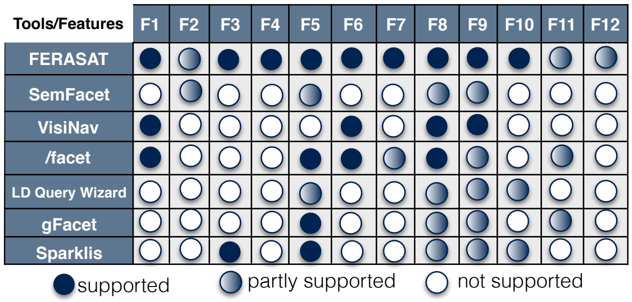
Figure 5: Employing the background knowledge to facilitate browsing of data.

showed the relevant characteristics of the projects, such as organizations involved, the organization type, and the program the project belongs to. The CORDIS dataset contains among others a text summarizing the content of the projects. Using the annotator tool helped me to extract general encyclopedic concepts from these textual descriptions and enabled me to browse data using two new facets (F_7), one for extracted terms and one for categories these terms belong to (see Figure 5). Combining extracted terms has a great advantage, as we can combine technical research terms and policy related terms to retrieve the relevant projects (F_1). This may solve the problem of finding how research links to the grand societal challenges. This is a core problem in assessing relevance of research (described in technical terms and policy related terms). Because the resulting set for a very specific topic is generally not too large, I could even manually inspect the policy-science link. As an example, I looked at chemical research in H2020 projects, related to one of the societal challenges. There are quite some water related topics in the H2020 projects. In total 22.5% of the water projects seem related to chemistry. Going a little deeper into this case shows the multidisciplinary character of the water related research in H2020, and what disciplines are more and what are less important in this portfolio (F_5, F_8, F_9).

6 RELATED WORK

Path-finding on semantic graphs such as RDF graphs, where semantics of the relations between resources are explicitly defined, leads to discovering meaningful and insightful connections between multiple resources. That is the reason why most of the current research work which investigates serendipity on Linked Data is focused on novel approaches for semantic traversal of RDF graphs and thereby serendipitous discovery of new related nodes.

Tools such as Everything_Is_Connected_Engine[8] and DBpedia RelFinder[15] allow serendipitous *storytelling* and *relation extraction* which benefit from path-finding on general knowledge graphs. In addition to that, domain-specific knowledge graphs enable experts to reveal *unsuspected connections* and/or *hidden analogies*. For instance, the Linked Data version of the TCGA (The Cancer Genome Atlas Database)[28] allows bio-medical experts to discover how cancer types tend to metastasize into other cancer types and to serendipitously explore linked data to see how the rheology of certain cancer types affects this metastasis. Furthermore, serendipitous recommendation realized by LOD paths-based techniques has been incorporated into the design of many personalized systems to minimize blind spots in information delivery. For example, in [21], a serendipity-powered TV recommender using BBC programs dataset is presented.



Tools/Features	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
FERASAT	●	◐	●	●	●	●	●	●	●	●	◐	◐
SemFacet	○	◐	○	○	◐	○	○	◐	◐	○	○	○
VisiNav	●	○	○	○	○	●	○	●	●	○	○	○
/facet	●	○	○	○	●	●	◐	●	◐	○	◐	○
LD Query Wizard	○	○	○	○	◐	○	○	◐	◐	◐	○	○
gFacet	○	○	○	○	●	○	○	◐	◐	○	◐	○
Sparklis	○	○	●	○	●	○	○	◐	◐	◐	○	○

● supported ◐ partly supported ○ not supported

Figure 6: Comparing existing RDF facted browsers based on the proposed serendipity design features.

There are also several tools and related works in the area of Linked Data-based faceted browsing which do not claim explicitly for a contribution in terms of serendipity (cf. Figure 6):

SemFacet [3] is a faceted search tool enhanced by the Semantic Web technologies to allow browsing of interlinked documents. SemFacet is implemented on top of a fragment of Yago and DBpedia abstracts. On contrary to FERASAT that focuses its results on a specific set of resource types, SemFacet allows refocusing results which could be implemented as an enhancement in the FERASAT environment to further support F_8. Although SemFacet exploits ontology-based reasoning for generation of facets and queries, no user-contextualization is supported. The main advantages of FERASAT over SemFacet are supports for customized interactive facet visualizations and enabling federated SPARQL queries over multiple knowledge graphs tailored based on the user context. *VisiNav* [12] is another linked data navigation system which combines features such as keyword search, object focus, path traversal and facet selection to browse web of data with a large variance. Although VisiNav provides some mechanisms to address the issue of *naked objects* (i.e. objects that are displayed without type-specific styling), it does not provide any personalized integrated view on distributed knowledge graphs. In our opinion, VisiNav acts more as a tool for traversing web of data rather than direct knowledge discovery tool. *\facet* [15] is a linked data faceted browser very similar to FERASAT but with limited capabilities to share the generated queries, adapt the results based on user context and to invert and randomize the facets for increasing the chance encounter. \facet enables multi-type browsing experience and allows adapting the dynamically generated facets based on their RDF relations. It also allows users to create facet specifications and build facet dependent visualizations and interactions to make surprising observations more noticeable. *Linked Data Query Wizard* [16] is a linked data browsing UI, heavily dependent on RDF Data Cube standard, which turns graph-based data into a tabular interface with supports for search and filtering to facilitate exploring linked data. Althought converting graphs to interactive spreadsheet tables increases the

[15] http://www.visualdataweb.org/relfinder.php

learnability of UI for users, it also results in limited capabilities for serendipity by limiting the flexibility of information visualization related to certain dimensions of data. *gFacet* [13] is a graph-based faceted browser which allows users to build their facets of interest on the fly. It enable users to perform a pivot operation and switch a facet to a result list. Color coded facets and their relationships facilitate explaining the surprising observations. However, no mechanism for sharing the query results, inversting and randomizing values is offered. *Sparklis* [10] is a query-based faceted search UI that uses the expressivity of natural language to facilitate browsing Linked Data and understaning the generated query. It does not exploit any interactive visualizations in the facets to make surprising observations more noticeable. Also using only a single facet on a single knowledge graph, to browse data, makes the divergent information behaviour difficult to achieve, though it increases the expressiveness and scalability of the traversed paths.

To the best of our knowledge, the related work in the domain of Linked Open Data where other aspects of serendipity than mere semantic path-finding are addressed, is quite scarce. The closest to our work is [9] which is based on SPARQL querying perspective where authors propose a query modification process to support serendipity features F_{11}: analogy, F_1: surprising observation, F_3: inversion, and F_4: disturbance. What distinguishes our approach from the above work is our more comprehensive investigation of serendipity design features and their implications on linked data faceted browsing environments for fostering serendipity on LOD.

7 CONCLUSIONS AND FUTURE WORK

Linked Open Data provides a rich domain for people to experience serendipity — finding valuable or agreeable things initially not sought for. Serendipity is a by-catch, an outcome or a moment of successful retrieval when a user is browsing data. In this paper we presented a set of serendipity-fostering design features amenable to data-driven systems together with a set of UI and Semantic Web techniques which support those features when a user is exploring linked data on a faceted browsing environment. To showcase the applicability of our proposal, we implemented a data-aware faceted browser UI to foster accidental knowledge discovery while browsing data scattered over multiple knowledge graphs.

As future work, we plan to implement more serendipity-fostering strategies within our faceted browser environment, in particular for detecting successful errors and performing cross-domain analogical reasoning. We also envisage to evaluate the usability of our implementation using a rigorous evaluation framework and also to extend its application to other domains such as life sciences.

Acknowledgement. This work was supported by a grant from the European Union's 7th Framework Programme provided for the project RISIS (GA no. 313082).

REFERENCES

[1] A. Acosta. Using serendipity to advance knowledge building activities. *Ontario Institute for Studies in Education, University of Toronto, Canada*, 2012.

[2] G. Alemu, B. Stevens, P. Ross, and J. Chandler. Linked data for libraries: Benefits of a conceptual shift from library-specific record structures to rdf-based data models. *New Library World*, 113(11/12):549–570, 2012.

[3] M. Arenas, B. C. Grau, E. Kharlamov, S. Marciuska, D. Zheleznyakov, and E. Jiménez-Ruiz. Semfacet: semantic faceted search over yago. In *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, pages 123–126, 2014.

[4] L. Björneborn. Design dimensions enabling divergent behaviour across physical, digital, and social library interfaces. In *PERSUASIVE*. Springer, 2010.

[5] F. Bobillo, M. Delgado, and J. Gómez-Romero. An ontology design pattern for representing relevance in owl. *The Semantic Web*, pages 72–85, 2007.

[6] P. H. Cleverley and S. Burnett. Retrieving haystacks: a data driven information needs model for faceted search. *Journal of Information Science*, 41(1), 2015.

[7] M. Cunha. Serendipity: Why some organizations are luckier than others. *Universidade Nova de Lisboa (Ed.), FEUNL Working Paper Series*, 2005.

[8] L. De Vocht, S. Coppens, R. Verborgh, M. Vander Sande, E. Mannens, and R. Van de Walle. Discovering meaningful connections between resources in the web of data. In *LDOW*, 2013.

[9] J. S. A. Eichler, M. A. Casanova, A. L. Furtado, L. A. P. P. L. Lívia Ruback, G. R. Lopes, B. P. Nunes, A. Raffaetà, and C. Renso. Searching linked data with a twist of serendipity. In *CAISE2017*, 2017.

[10] S. Ferré. Expressive and scalable query-based faceted search over sparql endpoints. In *International Semantic Web Conference*, pages 438–453. Springer, 2014.

[11] E. Hargrave-Thomas, B. Yu, and J. Reynisson. The effect of serendipity in drug discovery and development. *Chemistry in New Zealand*, 2012.

[12] A. Harth. Visinav: A system for visual search and navigation on web data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(4):348–354, 2010.

[13] P. Heim, T. Ertl, and J. Ziegler. Facet graphs: Complex semantic querying made easy. *The Semantic Web: Research and Applications*, pages 288–302, 2010.

[14] J. Heinström. Psychological factors behind incidental information acquisition. *Library & Information Science Research*, 28(4):579–594, 2007.

[15] M. Hildebrand, J. van Ossenbruggen, and L. Hardman. /facet: A browser for heterogeneous semantic web repositories. *The Semantic Web-ISWC 2006*, pages 272–285, 2006.

[16] P. Hoefler, M. Granitzer, E. E. Veas, and C. Seifert. Linked data query wizard: A novel interface for accessing sparql endpoints. In *LDOW*, 2014.

[17] D. Karger and M. Schraefel. The pathetic fallacy of rdf. Position Paper for SWUI06, 2006.

[18] A. Khalili and K. A. de Graaf. Linked data reactor: Towards data-aware user interfaces. In *Proceedings of the 13th International Conference on Semantic Systems*, SEMANTiCS 2017. ACM, 2017.

[19] A. Khalili, A. Loizou, and F. van Harmelen. Adaptive linked data-driven web components: Building flexible and reusable semantic web interfaces. In *ESWC2016*, pages 677–692, 2016.

[20] H. Liu, Y. Wu, and YimingYang. Analogical inference for multi-relational embeddings. https://arxiv.org/abs/1705.02426, 2017.

[21] V. Maccatrozzo, M. Terstall, L. Aroyo, and G. Schreiber. Sirup: Serendipity in recommendations via user perceptions. In *IUI2017*, pages 35–44. ACM, 2017.

[22] G. Marchionini. Exploratory search: from finding to understanding. *Commun. ACM*, 49(4):41–46, Apr. 2006.

[23] R. Merton and E. Barber. *The Travels and Adventures of Serendipity: A Study in Sociological Semantics and the Sociology of Science.* Princeton Press, 2004.

[24] D. A. Norman. *Living with complexity.* MIT press, 2010.

[25] D. A. Norman. *The Design of Everyday Things: Revised and Expanded Edition.* Basic Books, Inc., New York, NY, USA, 2013.

[26] E. Oren, R. Delbru, and S. Decker. Extending faceted navigation for rdf data. In *International semantic web conference*, volume 4273, pages 559–572. Springer, 2006.

[27] N. Ramakrishnan and A. Y. Grama. Data mining: From serendipity to science. *Computer*, 32(8):34–37, 1999.

[28] M. Saleem, M. R. Kamdar, A. Iqbal, S. Sampath, H. F. Deus, and A.-C. Ngonga. Fostering serendipity through big linked data. *ISWC SemWeb Challenge*, 2013.

[29] N. Shadbolt, T. Berners-Lee, and W. Hall. The semantic web revisited. *IEEE intelligent systems*, 21(3):96–101, 2006.

[30] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *AVI2006 workshop*, pages 1–7. ACM, 2006.

[31] V. C. Storey, V. Sugumaran, and A. Burton-Jones. The role of user profiles in context-aware query processing for the semantic web. In *Intl. Conference on Application of Natural Language to Information Systems*, pages 51–63. Springer, 2004.

[32] E. G. Toms et al. Serendipitous information retrieval. In *DELOS Workshop*, pages 17–20, 2000.

[33] P. van Andel. Anatomy of the unsought finding. serendipity: Origin, history, domains, traditions, appearances, patterns and programmability. *Br J Philos Sci*, 45(2):631–648, June 1994.

[34] P. van Andel and D. Bourcier. Serendipity & abduction in proofs, presumptions & emerging laws. In *The Dynamics of Judicial Proof*, pages 273–286. Springer, 2002.

[35] L. Yu. *Follow Your Nose: A Basic Semantic Web Agent*, pages 711–736. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.